

In each video, the left side shows the original textual description and corresponding motion sequence, while the right side presents the decomposed sub-texts generated by ChatGPT and the motion segments obtained through text-motion joint segmentation. The analysis is provided below.

Example 1	Correct.
Example 2	Correct.
Example 3	Correct.
Example 4	The last segment (“touch something”) partially includes the semantics of “step around to their right.”
Example 5	The last segment (“halts and sits down”) partially includes the semantics of “walks three steps forward.”